Machine Learning
Lab Assignment 7
(Decision Trees)

Note: This assignment is for your understanding of the application of decision trees for classification and regression.

Question 1: Classification

Consider the CarEvaluation dataset (Available from: `https://archive.ics.uci.edu/dataset/19/car+evaluation`) and implement the following:

- Grow the decision tree using the entropy measure (ID3 algorithm) or gini index (CART algorithm). Decision trees tend to be overly complex and do not generalize well to the data, leading to overfitting. To overcome this, use **decision tree pruning** to address overfitting. Prune the tree using cost complexity pruning (also known as weakest link pruning) and use cross-validation to find the optimal value of the complexity parameter α. Plot the average accuracy vs. α. Visualize the pruned decision tree constructed on the training data.
- Use the inbuilt scikit-learn implementation for the above experiments.

Question 2: Regression

On the Boston housing dataset (`http://lib.stat.cmu.edu/datasets/boston`), perform decision tree regression to predict the value of `MEDV`, representing the median price of owner-occupied homes.

- Grow the regression tree using recursive binary splitting. Decision trees tend to be overly complex and do not generalize well to the data, leading to overfitting. To overcome this, use **decision tree pruning** to address overfitting. Prune the tree using cost complexity pruning and use cross-validation to find the optimal value of the complexity parameter α. Visualize the pruned decision tree constructed on the training data.
- Plot the regression curve for `MEDV` vs. `CRIM` (per capita crime rate by town), `MEDV` vs. `INDUS` (proportion of non-retail business acres per town), and `MEDV` vs. `AGE` (proportion of owner-occupied units built prior to 1940).
- Use the inbuilt scikit-learn implementation for the above experiments.